More Than the Sum of Its Parts

Tak, Lineke M.; Meijer, Anna; Manoharan, Andiappan; de Jonge, Peter; Rosmalen, Judith G. M.

Published in:
Psychosomatic Medicine

DOI:
10.1097/PSY.0b013e3181d714e1

Publication date:
2010

More Than the Sum of Its Parts: Meta-Analysis and Its Potential to Discover Sources of Heterogeneity in Psychosomatic Medicine

Lineke M. Tak, MD, PhD, Anna Meijer, MSc, Andiappan Manoharan, MSc, Peter de Jonge, PhD, and Judith G. M. Rosmalen, PhD

Meta-analyses may contribute to more reliable knowledge about the existence of certain relations in the area of psychosomatic research. Surprisingly, the increasing popularity of meta-analysis is not reflected in the number of meta-analyses of observational studies published in *Psychosomatic Medicine*. This may be due to the specific difficulties that apply to meta-analyses of observational research. The aim of this paper is to provide a nontechnical overview of the principles of meta-analysis applied to observational research. We will highlight general principles of meta-analysis and discuss the major threats to its validity, with an emphasis on its specific merits and pitfalls for psychosomatic research, using several examples. We conclude that meta-analysis is a relatively simple technique, leaving little reason for not routinely applying it when performing a systematic review. An adequately conducted meta-analysis may not only provide a summary estimate of a certain association, but it has additional value in discovering relevant confounders, mediators, and moderators, as well as identifying areas of research that require more attention.
Key words: meta-analysis, effect size, heterogeneity, meta-regression, mega-analysis, observational studies.

BMI = body mass index; **IL** = interleukin; **IPD** = individual patient data; **OR** = odds ratio; **PTSD** = posttraumatic stress disorder; **SEM** = standard error of the mean; **SMD** = standardized mean difference.

INTRODUCTION

Meta-analysis is a statistical procedure that integrates several studies concerning a certain research question to reach a more secure conclusion. State-of-the-art meta-analyses have the potential to provide a more objective appraisal of the evidence than traditional narrative reviews. They can reveal that repeated results in the same direction across several studies, even if not one is significant, can be much more powerful evidence than a single significant result from an individual study (1). In addition, meta-analyses can provide insight into why different studies have found different results. Furthermore, whereas adequately powered, randomized, controlled trials have a relatively high positive predictive value of reflecting the true relationship, this value drops for research findings of observational studies (2). Meta-analysis of reported associations in observational studies may raise the positive predictive value of the true relationship. Finally, meta-analysis may identify areas of research that need more investigation, for example, based on results of subgroup or sensitivity analyses (3,4).

Meta-analysis is a particularly useful procedure in psychosomatic research. Many studies in this field meet characteristics that contribute to the risk of nonreplication, such as having small sample sizes, retrieving small effect sizes, testing a large number of relationships without clear rationale, and having large flexibility in designs, definitions, outcomes, and analytical methods (2,5). In spite of this, the sharply increased number of published meta-analyses in general medicine (6,7) is not reflected in *Psychosomatic Medicine*, where the average number of meta-analyses published is stable at around two per year (Table 1). Since 2000, 15 of 16 meta-analyses published in *Psychosomatic Medicine* were purely based on observational research (and the other one was a combination of observational and intervention research) (8), whereas this was the case for only one fourth of the 127 meta-analyses published in this time period in the *Journal of the American Medical Association* as an example of a general medical journal. In the light of certain difficulties that are associated with meta-analyses of observational studies, most importantly the risk to produce very precise but equally spurious results (9), the relative low number of meta-analyses in *Psychosomatic Medicine* may not be surprising.

Since the last review on meta-analysis in *Psychosomatic Medicine* almost two decades ago (10), new techniques, procedures, and recommendations have become available (6). The current paper provides an update in the form of a nontechnical overview of the general principles of meta-analysis to readers with a basic knowledge of statistics. We will highlight the general principles of meta-analysis and discuss the major threats to its validity, with an emphasis on its specific merits and pitfalls for observational, psychosomatic research. Our aim is that, after reading our paper, researchers should be able to critically interpret meta-analyses performed by others or are encouraged to perform a meta-analysis themselves.

Inclusion of Studies

Literature Search

In searching literature for a meta-analysis, the goal is to include as many of the existing relevant studies as possible in a reproducible manner. Meta-analyses on psychosomatic research should use at least the literature databases Medline and Embase. In addition, other, more subject-specific databases can be searched, depending on the research question (e.g., PsycINFO, CENTRAL, CINAHL, ISI Science and Social Science Citation Index, Cochrane Library). Each database has

From the Interdisciplinary Center for Psychiatric Epidemiology (L.M.T., A.M., P.d.J., J.G.M.R.), University Medical Center Groningen, University of Groningen, Netherlands; and the Department of Biostatistics and Computing (A.M.), Institute of Psychiatry, King's College London, United Kingdom.

Address correspondence and reprint requests to Judith G. M. Rosmalen, PhD, Interdisciplinary Center for Psychiatric Epidemiology, CC72, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700 RB, Groningen, Netherlands. E-mail: J.G.M.Rosmalen@med.umcg.nl

Received for publication May 11, 2009; revision received November 11, 2009.

Dr. de Jonge is supported by VIDI Grant 016.086.397 from the Dutch Medical Research Council. Dr. Rosmalen is supported by VENI Grant 016.056.064 from the Dutch Medical Research Council.

The authors have not disclosed any potential conflicts of interest.

DOI: 10.1097/PSY.0b013e3181d714e1

0033-3174/10/7203-0253

TABLE 1. Overview of Meta-Analyses Published in *Psychosomatic Medicine* Since 2000

First Author	Year	Association	*n* of Primary Studies	Effect Size Measure	Fixed or Random Effects	Homo- or Heterogeneity	Moderator Analyses	Meta-Regression	Quality Assessed	Publication Bias
Howren (55)	2009	Depression and inflammation	9–61	*d*	Random	Heterogeneity	5–10	No	−	Funnel plot, fail safe *N*
Chida (62)	2008	Positive well-being and mortality	19–21	HR	Random	Heterogeneity	5–10	No	+	Funnel plot, Egger's test, fail safe *N*
Chida (76)	2008	Psychosocial factors and atopy	9–34	*r*	Random	Heterogeneity/ homogeneity	5–10	No	+	Begg's test
Steffen (77)	2006	Acculturation and blood pressure	114–124	*d*	Random	Heterogeneity	5–10	No	+	Funnel plot, fail safe *N*
Carter (78)	2006	Cesarean section and postpartum depression	8	OR	Not stated	Heterogeneity	0	No	+	Not assessed
Cho (8)	2005	Placebo response in chronic fatigue syndrome	29	Proportion	Random	Heterogeneity	<5	Yes	+	Not assessed
Barth (39)	2004	Depression in coronary heart disease and mortality	4–7	OR, HR	Random	Homogeneity/ heterogeneity	<5	No	−	Funnel plot
Van Melle (38)	2004	Post-MI depression and cardiovascular prognosis	6–9	OR	Fixed/random	Homogeneity/ heterogeneity	<5	No	−	Funnel plot, Egger's test
Dickens (28)	2003	Depression and pain perception	2–6	*d*	Fixed	Homogeneity	0	No	−	Funnel plot
Sundin (79)	2003	Major life events and stress by Horowitz Event Scale	66	n.a.	n.a.	n.a.	n.a.	Yes	−	Not assessed
Wulsin (40)	2003	Depressive symptoms and coronary heart disease	10	OR	Random	Heterogeneity	0	No	+	Funnel plot, Begg's test, file drawer *N*
Henningsen (80)	2003	Depression and anxiety in functional syndromes	3–45	*d*	Fixed/random	Homogeneity/ heterogeneity	<5	No	−	Fail safe *N*
Dickens (81)	2002	Rheumatoid arthritis and depression	9–12	*r*	Not stated	Heterogeneity	5–10	No	+	Fail safe *N*, file drawer *N*
Rutledge (30)	2002	Psychological factors and hypertension	15	*r*	Random	Homogeneity	5–10	No	+	Fail safe *N*
De Groot (82)	2001	Depression and complications of diabetes	22	*r*	Random	Heterogeneity	<5	No	−	Fail safe *N*
Buckley (31)	2001	Cardiovascular measures in PTSD	4–28	*d*	Not stated	Homogeneity	<5	No	−	Not assessed

HR = hazard ratio; OR = odds ratio; MI = myocardial infarction; n.a. = not applicable; PTSD = posttraumatic stress disorder.

specific search possibilities, and most databases provide tutorials. Databases often use key words, but free text should also be searched (11,12). The search should be conducted without language restrictions to reduce the risk of language bias (13,14). Because searching the literature is almost a specialty in itself and errors in search strategies are common (15), additional consultation of a librarian may be worth considering. The probability that research findings are published is influenced by the nature and the direction of the results. Significant research findings are overrepresented, whereas results conflicting with the prevailing beliefs about the association are underrepresented (16). Searching for unpublished studies is thus important to achieve a representative sample of the work available in the research area under study, but it requires considerable effort. Unpublished findings may, for example, be revealed by asking relevant research groups for any unpublished results or checking dissertation databases (e.g., Dissertation Abstracts Online, ProQuest Dissertations, and Theses).

Selection of Relevant Studies

Articles are selected for inclusion based on a predesigned protocol containing inclusion criteria specifying the type of subjects, exposure, outcomes, and type of study (11,12,17). This is preferably done by two independent reviewers, as they select on average 9% more studies than one (18). One particular problem in the selection process is the fact that several articles with different first authors may report on the same study, or on partly overlapping data. This problem may especially occur in observational studies that gather information on a large number of variables over a relatively long period of time, resulting in more than one publication on a single study. Just like the search strategy, the selection process should be reported in detail.

Methodological Quality

Critical appraisal of the methodological quality of primary studies is an essential feature of meta-analysis. Good methodological quality can be defined as having a design that minimizes bias in the estimation of the association under study. Critical appraisal checklists or scales ("tools") can be used as a threshold for inclusion of studies, or preferably, the meta-analysis can be repeated excluding low-quality studies to assess whether results would change. Although there is a plethora of tools for assessing quality of intervention trials, consensus on the ideal tool for assessing methodological quality of observational studies is currently not available (19). Major domains that should be incorporated in every observational studies quality tool are selection of participants, measurement of dependent variables, and control for confounding.

The type of tool used to assess quality can dramatically influence the interpretation of meta-analyses (20). To develop a valid tool, experts in the field could be consulted and development of the tool should be clearly stated. Reliability of the tool can be assessed by using at least two independent raters to score the individual papers, and interrater agreement statistics should be reported. Researchers should be aware that using a quality tool inevitably introduces subjectivity, such as the decisions on which items to include and on the scoring rules for each quality item. When developing a quality tool, general items for quality of reporting can be used, such as consensus guidelines on reporting of randomized trials, CONSORT (21); diagnostic tests, STARD (22); or observational studies in epidemiology, STROBE (23). Recommendations (24) for adequate reporting of case-control studies in the psychiatric setting have also been made, which are largely applicable to observational psychosomatic research. Additionally, researchers can include specific items that are pivotal for good quality studies in their field. Assessing sources of bias is a crucial but equally complex function of a quality tool, because distinguishing quality of reporting and quality of the actual study design is often not possible. Notwithstanding some degree of uncertainty about the validity of comparing study quality, quality tools specifically designed for a meta-analysis of a certain topic under study may additionally serve as a guideline for conducting high-quality future research. An example of a quality tool for meta-analysis of observational studies in the psychosomatic area is one developed for studies on cardiac vagal activity in functional somatic syndromes. Experts in the field were asked to review this quality tool that includes items such as whether the functional somatic syndrome has been reliably assessed; whether the measurement of cardiac vagal activity has been reported in appropriate unit, and whether specific covariates, such as age, gender, body mass index (BMI), depression, and medication use, have been assessed or adjusted for (25). In this meta-analysis, it could not be proven that study quality accounted for the mixed findings. It was advised, however, that future research adhering to the proposed quality criteria may provide a more definite answer on the question whether lower cardiac vagal activity is involved in the etiology of functional somatic syndromes. The assessment of methodological quality should be considered as a routine procedure in meta-analysis.

Meta-Analysis Performance

Effect Size per Study

The basic information needed for a meta-analysis is the effect size per study, which is the measure of the magnitude (strength) of the association between two variables. Information needed to calculate this effect size consists of a summary measure and a measure of its precision (standard error or 95% confidence interval). Widely used summary measures are the correlation coefficient, odds ratio (OR), and standardized mean difference (SMD), but mean difference, risk ratio, rate ratio, hazard ratio, proportion, etc. are also possible summary measures. In case of variability in reported effects, several formulas for converting test statistics (such as t, χ^2, Z, or F values, or their associated p levels) to effect size estimates (such as Cohen's d, OR, and correlation coefficients) and

formulas for converting effect size estimators from one type to another are available (26,27).

In the area of psychosomatic research, however, there may be specific problems concerning retrieving effect sizes, as different measures for the same construct are often applied in the original studies. For example, in a meta-analysis on depression and pain perception thresholds, effect sizes had to be calculated from studies using different methods to measure depression (Beck Depression Inventory, Hamilton Depression Scale, or diagnostic criteria according to *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*) and pain perception (cold-, heat-, pressor-, ischemic-, and electrical stimuli) (28). Effect sizes based on the standard deviation (SD), such as the SMD, provide a solution to this problem. The SMD is calculated as the difference between the means of the cases and controls divided by the pooled SD. Although the SMD is the predominantly used effect size in social sciences (1), it has been disputed by others (29). The use of SMDs is criticized because of the underlying assumption that the different scales from which the SMD is calculated are linearly related. Although this is a limitation of the SMD, calculating SMDs is sometimes the only feasible option (e.g., in cross-sectional case-control studies).

Correlation Coefficients and SMDs

When an effect size is applied to continuous variables, commonly used effect size indices are the *r* family and the *d* family (1). The *r* effect size family includes all types of correlation coefficients (i.e., *r, phi, rho*) and is preferably used when the studies comprising the meta-analysis primarily report the correlation between variables (e.g., continuously measured psychological factors, such as anger or hostility, and development of hypertension in population cohorts) (30). The *d* effect size family provides SMDs and is preferably used when the studies comprising the meta-analysis primarily report analyses of variance and *t* test comparisons between groups on continuous variables (e.g., cardiovascular activity in posttraumatic stress disorder [PTSD] patients versus healthy controls) (31). Cohen's *d* and Hedges's *g* are two widely used examples of an SMD. When sample sizes are small, Cohen's *d* may produce estimates of the population effect size that are slightly too large (32). Therefore, Cohen's *d* is sometimes adjusted by the following formula ($\sqrt{}$ [(*n* cases + *n* controls − 2)/(*n* cases + *n* controls)] × *d*), which results in Hedges's *g* (33). For example, a study compared cardiac vagal activity in 11 patients with chronic fatigue syndrome (2.75 SD 1.39 ln ms^2) with 11 healthy controls (3.09 SD 0.56 ln ms^2) (34). It can be calculated that Cohen's $d = -0.34$ (95% confidence interval [CI], 0.00 to −1.14). In this small sample, correction leads to a small attenuation of the effect size, namely, Hedges's $g = \sqrt{}$ [(11 + 11–2)/(11 + 11)] × −0.34 = −0.32. The *r* of this association is .17 based on Cohen's d ($r = \sqrt{}$ [d^2)/]d^2) + 4), indicating that explained variance (r^2) chronic fatigue syndrome by lower cardiac vagal activity in this study is $.17^2 = 2.8\%$ (and 2.5% when using Hedges's *g*).

Occasionally, descriptive or inferential statistics needed to compute an effect size are not reported. Conservative approaches to impute an effect size for missing values exist. For example, when a significant association was reported in the primary study, a conservative effect size assuming that *p* was equivalent to .05 can be computed. In case there was no significant association, an effect size of 0.00 can be imputed. Several methods to impute missing data in meta-analyses have been discussed elsewhere (35).

The magnitude of effect sizes is often interpreted by using Cohen's conventions, in which an SMD of 0.0 means no difference, 0.2 represents a small difference, 0.5 represents a moderate difference, and 0.8 represents a large difference (36). Inherent to the multifactorially caused conditions typically under investigation in psychosomatic research, effect sizes are usually small. For example, the median Cohen's *d* in the studies listed in Table 1 is 0.34 (interquartile range, 0.27–0.55). The final evaluation of the meaning of the effect size, nevertheless, requires individual judgment regarding the specific topic under study, in which the consequence of the outcome or the possibility of prevention and treatment are also taken into account.

ORs and Other Probability Effect Sizes

Probability effect sizes are usually given in studies with a binary outcome, such as in studies with disease versus no disease or mortality versus no mortality as end point. The selection of the appropriate summary statistic is a subject of debate due to conflicts in the relative importance of mathematical properties and the ability to interpret results intuitively (6). Recommendations (37) on how to choose between ORs, risk differences, risk ratios, and other relative measures have been documented elsewhere. Frequently, ORs are reported, such as in a meta-analysis on the association between depression and cardiovascular disease and mortality (38–40). The odds are the number of patients who fulfill the criteria for a given end point divided by the number of patients who do not. The OR relatively easily allows combining data and testing for statistical significance. Although relative effect measures are generally used for summarizing the evidence, absolute measures, such as the absolute risk reduction or the number of patients needed to treat to prevent one event, are more useful when applying the results in a concrete clinical or public health situation and should be recalculated from the relative summary estimates (4,41). Also, effect sizes should be compared with the effects of well-established risk factors in the field to determine their (clinical) importance. For example, the influence of depression as a risk factor for the development of coronary disease in community samples without clinically apparent heart disease was considered similar to published effect sizes of the widely accepted risk factor smoking (40).

META-ANALYSIS IN PSYCHOSOMATIC MEDICINE

Pooling of Effect Sizes

Methods for calculating the summary estimate by combining the effect sizes of the individual studies use a weighted average of the results, in which larger studies have more influence than smaller ones. This study weight is computed from the variance or squared standard error of the mean (SEM): weight factor (w) = $1/\text{SEM}^2$. The summary estimate in a meta-analysis is the mean weighted effect size, calculated by the sum of the products of effect size and weight per study, divided by the sum of all study weights (Σ effect size $\times$ $w/\Sigma w$). The accompanying 95% CI can be calculated with the following formula: $\pm 1.96 \sqrt{}$ 1/[SEM of summary estimate]^2. It is assumed that each value contributing to the summary estimate is statistically independent of the others. An extensive overview of the statistical basis of formal meta-analysis has been provided by others (42). The meta-analysis can be repeated using different methods to assess whether the same results are achieved and the summary estimate is robust to the decisions made to obtain it.

Software

Programs designed for the statistical pooling of data in meta-analysis are available, and most general statistical packages include meta-analysis options. Most of these programs are relatively easy to master and offer tutorials and a help function. Moreover, many programs and add-ons to statistical packages are freely available on the Internet. In general, programs offer at least basic statistical methods and graphical presentations, and commercial software is not necessarily better than free software (43). Differences may exist in statistical methods, usability, graphics, and whether or not the software is being maintained. Some of the most-used programs will be discussed and links to more information will be provided (Table 2). The basic results obtained from the different software packages are essentially the same (44). The studies of Bax et al. (43) and Egger et al. (44) provide an overview and comparison of some meta-analysis programs.

Fixed Effect Versus Random Effects Models

When pooling the effects of all studies included in the meta-analysis, the fixed effect model or the random effects model can be used. The fixed effect model assumes that the samples of all studies are based on the same population and that the same underlying effect is thus measured in all studies (i.e., there is one true effect size). In this method, between-study variation is assumed to be due to sampling error. A disadvantage of the fixed effect model is that it is highly unlikely that studies do measure the same underlying effect, especially in epidemiologic research (45). The random effects model, in contrast, assumes that each sample comes from a

TABLE 2. Overview of Software Packages for Meta-Analysis

Software	Availability	Web Site
Arcus Quickstat	Commercial	http://www.camcode.com./
Comprehensive meta-analysis	Commercial	http://www.meta-analysis.com
DSTAT	Commercial	http://johnson.socialpsychology.org/ Johnson BT, editor. DSTAT 1.10: Software for the Meta-Analytic Review of Research Literature. Hillsdale, NJ: Lawrence Erlbaum; 1993
Easy MA	Free	http://www.spc.univ-Lyon1.fr/~mcu/easyma/
Fast*Pro	Commercial	Eddy DM, Hasselblad V, Schachter R. Meta-Analysis by the Confidence Interval Method. The Statistical SynthesisoOf Evidence. San Diego: Academic Press; 1992
META	Commercial	http://userpage.fu-berlin.de/health/meta_e.htm Schwarzer R. Meta-Analysis Programs (Computer Program and Manual) v. 5.3. Berlin, West Germany: Institute fur Psychologie, Freie Universitat Berlin; 1989
Meta analysis easy to answer	Free	http://davidakenny.net/meta.htm
Meta Stat	Free	http://ericae.net/meta/metastat.htm
Meta-analysis	Commercial	Software accompanying book
Meta-analyst	Free	http://www.medepi.net/meta/MetaAnalyst.html
MetaWin	Commercial	http://www.metawinsoft.com/
MIX	Free	http://www.mix-for-meta-analysis.info/download/index.html
RevMan	Free	http://www.cc-ims.net/revman
SAS		www.sas.com
S-plus		www.splus.com
SPSS		www.spss.com
STATA		www.stata.com
True epistat		Epistat Services, Richardson, TX Gustafson TL. True Epistat Reference Manual. Richardson, TX: Epistat Services; 1994
WEasy MA	Commercial	www.clininfo.fr/uk/index.html Chevarier P, Cucherat M, Freiburger T, Maupas J, Visele N, Buguard F, Bazog P. WeasyMA. Lyon: ClinInfo; 2000

different population and that the effects in these populations may also differ. Between-study variation is assumed to be due to differences in the underlying effects in the samples. The random effects model gives the average effect of all studies (12). A disadvantage of the random effects model is that it assumes the studies are a random sample of effect sizes and that between-study variation is distributed normally (16). This is often not the case, for example, as a result of publication bias (45). An advantage of the random effects model is that it permits to generalize to studies that might be done in the future.

In the fixed effect model, the inverse variance method is used to pool effect sizes based on continuous data, such as mean differences or SMDs. To pool effect sizes based on binary data, such as ORs and relative risk, the Mantel-Haenszel's method can be used (46), or the Peto method in case of pooling ORs of studies with balanced arm sizes, small intervention effects, or rare events (47). When the effect sizes are pooled using a random effects model, the DerSimonian-Laird method is used both for effect sizes-based binary and continuous data (48).

The random effects model is more conservative than the fixed effect model and is used when heterogeneity is suspected (4). Although tests for heterogeneity are often used to determine whether a fixed or random effects model must be used, these tests are often underpowered, and deciding on the model should therefore be primarily based on characteristics of included studies (12,16). In general, the random effects model is more plausible, and using the fixed effect model should only be done when this can be firmly justified on theoretical grounds. An example of how using random effects versus fixed effect analysis can change the summary estimate and conclusion is found in a meta-analysis on cortisol levels in patients with functional somatic syndromes (49). The fixed effect model shows significantly lower cortisol in patients with functional somatic syndromes compared with healthy controls (SMD, −0.12; 95% CI, −0.18 to −0.05; $p < .01$), whereas the more appropriate random effects model shows a wider CI and no statistical significant difference (SMD, −0.07; 95% CI, −0.17 to 0.04; $p = .24$).

Forest Plot

The main results of a meta-analysis are usually represented in a forest plot. Forest plots graphically display information on the individual studies included in the meta-analysis, the amount of variation between studies, and an overall estimate of the results of all studies combined (Fig. 1) (50). ORs are best plotted on logarithmic scales, as this enables ORs of the same magnitude but opposite directions—for example, 0.1 and 10.0—to be equidistant from 1.0 (4). Next to the forest plot, the basic details of each study supplying data should be presented, such as primary author, year of study, design, crude data, derived summary estimate and measure of its precision, allowing readers to evaluate the summaries against what was presented in original reports, or to repeat the meta-analysis at the same time making other decisions or using other techniques.

Heterogeneity

Heterogeneity in meta-analysis means that included studies differ considerably on one or several important aspects, which may affect comparability of their results and which may have caused differences in results. Studies can be different in a) biological, psychological, or clinical variables, including gender, age, characteristics of study participants, severity of exposure, and condition or disease; b) methodological variables, including study design, measurement procedures, extent of control for confounding, and response measures; and c) miscellaneous variables, including year of publication, characteristics of the authors, and funding.

The presence of heterogeneity can be calculated statistically (51). The most used measures are the Q statistic, I^2, and tau-squared (τ^2). To begin with, the Q statistic (also called Cochran's χ^2 statistic) is a χ^2 test calculating whether variation in study results is due to chance or whether variation is due to systematic underlying differences and the null hypothesis should be rejected. A value of Q similar to the degrees of freedom in the analysis indicates little heterogeneity. When it is considerably higher, and the p is <.10, this indicates heterogeneity (52). The Q statistic, however, has a number of limitations. The Q statistic has low power when a single study largely contributes to the mean weighted effect size (16). It

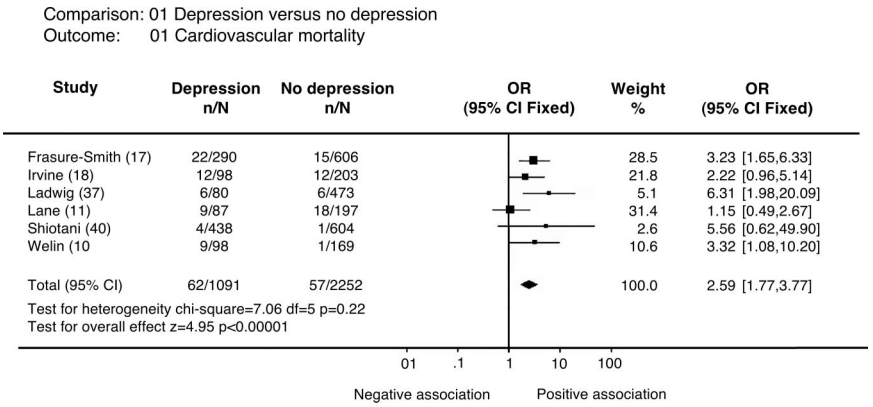

Figure 1. Example of a forest plot on the prognostic value of depression on cardiovascular mortality. Odds ratios are plotted on the logarithmic scale. *OR* = odds ratio; *CI* = confidence interval. Reprinted with permission from Van Melle et al (38).

also has low power when included studies are small or when there are few studies, whereas this test may detect heterogeneity even though it is not substantial when many studies are included (12). In the studies presented in Table 1, meta-analyses in which the Q test was not significant were usually based on a small number of primary studies (<10), whereas the Q test was significant in all meta-analyses with a relatively large number of included primary studies (>30). This implies that heterogeneity can be considered the rule, rather than the exception, in meta-analyses published in *Psychosomatic Medicine*. A second measure of heterogeneity is the I^2 statistic, which is a derivative of Q. This statistic gives the percentage of variability in results that is caused by heterogeneity rather than coincidence (12). Generally, an I^2 of $>50\%$ indicates considerable heterogeneity. A third statistic that is often used to report heterogeneity is τ^2. This is the variance of the true effect size; thus, there is no heterogeneity when this statistic is 0 (45). There are several more statistics that assess heterogeneity, some specific to the type of effect measure (16).

However, because of problems with power and accuracy, when statistical tests of heterogeneity indicate that the null hypothesis of no heterogeneity holds, this does not indisputably prove that studies are completely homogeneous (16). It should still always be investigated whether studies have important clinical and methodological differences. When heterogeneity is suspected, this must be accounted for in the statistical analysis and mentioned in the discussion. Usually, the presence of heterogeneity is considered to be a negative aspect of meta-analysis, because it makes the results of the meta-analysis difficult to interpret (16) and suggests that samples may be too different to be combined. However, as no widely accepted quantitative measure exists to grade heterogeneity, it may be better to examine it in a meta-analysis rather than use it as a reason for not conducting one (53). Heterogeneity can also have advantages. If studies that are clinically and methodologically heterogeneous lead to comparable results, this means that the results are generalizable to a wider population. In addition, investigating sources of heterogeneity can lead to a better understanding of associations, new hypotheses (11,16), and improvement of future research (54).

Moderator Analysis

Performing meta-analysis on subgroups based on characteristics that potentially are responsible for differences in effect sizes between studies can demonstrate whether the strength of the summary estimate is influenced by these characteristics. This procedure is referred to as moderator analysis. For example, in a meta-analysis, a significant difference was found in interleukin (IL)-6 serum levels between depressed patients and controls, with an SMD of 0.25, 95% CI, 0.18–0.31. However, the magnitude of the summary estimate of the association between IL-6 and depression was largely attenuated in studies that adjusted for BMI ($n = 22$; SMD, 0.08; 95% CI, 0.02–0.13; $p = .007$) as compared with studies without BMI adjustments ($n = 39$; SMD, 0.50; 95% CI, 0.37–0.63; $p < .001$) (55).

Ideally, such subgroup analyses are planned in advance, because investigating heterogeneity post hoc based on the data from the meta-analysis itself can lead to overinclusion (16). Providing a rationale for each moderator and giving due consideration to the role that each moderator is intended to play is essential (5). In case of post hoc subgroup analyses, results should be reported as exploratory and the need for replication should be mentioned (45). Particularly in observational studies, possible moderator analyses on confounders, moderators, and mediators are an important part of the meta-analysis. The extent to which putative confounders, moderators, and mediators have been taken into account in original studies is often highly variable, and extracting useful data is not always possible.

Some difficulties may arise when using terminology regarding confounders, moderators, and mediators. A variable may be considered as a confounding or mediating factor in the original study, but this variable is tested as a moderator in the meta-analysis. For example, authors in the previously mentioned meta-analysis on IL-6 and depression proposed that depressive symptoms may facilitate weight gain over time as a result of physical inactivity. In this pathway, BMI may be a mediator in reality (in case depression leads to weight gain and weight gain to inflammation), but it is referred to as a moderator of the effect size in the meta-analysis.

The difficulties faced in moderator analyses are many. First, there is the risk of spurious findings due to multiple testing. When the number of original studies in the meta-analysis is small (i.e., $n = 10–15$), there are insufficient degrees of freedom to test more than one moderator variable (56). Nevertheless, many more subgroup analyses are often performed, as illustrated by some of the studies listed in Table 1.

Second, when moderating variables are continuous, they have to be categorized to be able to perform a moderator analysis. It is, however, often unknown how to define subgroups. Artificially grouping data into categories introduces measurement error with an inevitable loss of power (5). Furthermore, the arbitrariness of the choice of cut point may lead to the undesirable temptation of trying more than one value and choosing the one that gives the most satisfactory result (41). Third, moderator analysis does not provide a statistical test of the existence of a moderator effect. Fourth, one cannot look at effect moderation at the same time keeping other covariates constant. When two moderators are highly correlated and the first causes changes in the effect size, a moderator test for the second will likely also be significant, even though this second moderator does not truly influence the strength of the effect.

An example of difficulties in interpreting the results of moderator analysis is a meta-analysis across 37 studies on cortisol levels in patients with PTSD (57). Overall, cortisol levels were not significantly different in PTSD patients compared with controls (SMD, −0.12; 95% CI, −0.32 to 0.08l

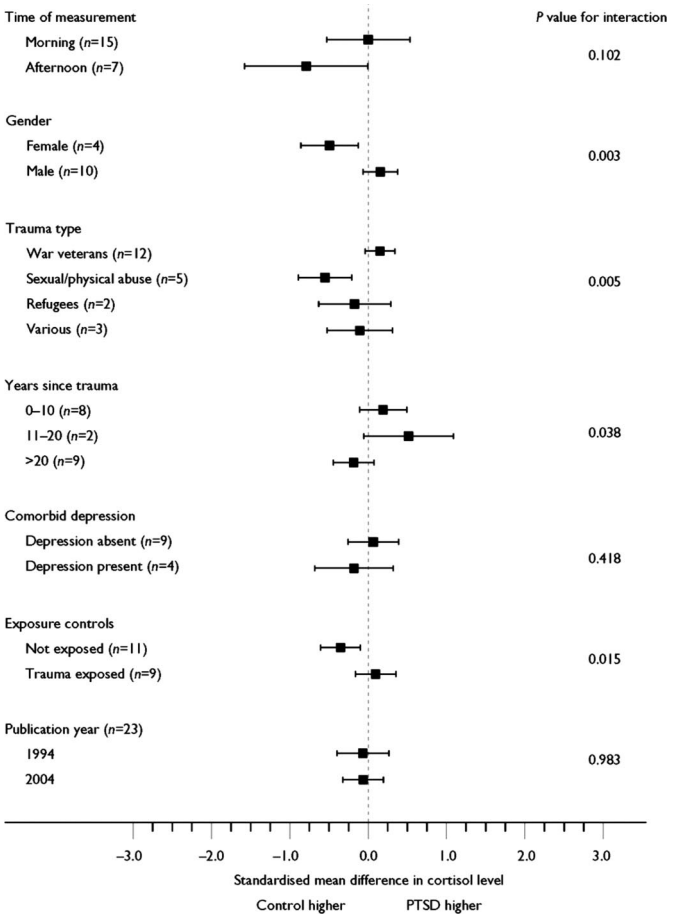

Figure 2. Examples of moderator analyses in a meta-analysis on cortisol and posttraumatic stress disorder (57). *PTSD* = posttraumatic stress disorder. Reprinted with permission from the Royal College of Psychiatrists (57).

$p = .24$) (57). Figure 2 presents subgroup analyses (SMD and 95% CI are shown) based on several moderators. Although there is no significant difference in cortisol between PTSD patients and controls in the primary summary estimates, significantly lower levels of cortisol are found in females with PTSD compared with female controls (gender is a moderator), in patients with PTSD with physical or sexual abuse compared with controls (trauma type is a moderator), and in PTSD patients when they are compared with controls without trauma exposure, as opposed to controls with trauma exposure but without PTSD (exposure status of control group is a moderator). However, as the authors mentioned, it is not possible to disentangle whether those moderators act separately from each other. For example, the association between female gender and lower cortisol could be explained by a larger prevalence of sexual abuse in women. In this case, meta-regression could be a possible solution to further elucidate the independent contribution of those factors.

Meta-Regression

Meta-regression is a regression-based analysis that aims to test for study heterogeneity by associating study characteristics with study outcome. Typically, the independent variables (predictors) are characteristics of each study, such as participants' mean age, proportion of women, or follow-up duration. The dependent (outcome) variable is the study effect size, such as the SMD or log OR. The procedure for multivariable meta-regression closely follows conventional regression analysis, the only difference being that a variable equal to the inverse variance (i.e., the study weight) has to be used as case weight to perform a weighted regression. Meta-regression can be used to explain heterogeneity and provides the possibility to simultaneously assess multiple characteristics. Again, the fixed effect or random effects model can be used for meta-regression. The full range of regression models and methods (i.e., linear or logistic regression, testing interactions, model fitting statistics) can be employed (29). For example, in a meta-analysis on placebo response in patients with chronic fatigue, it was found that the placebo response was higher in interventions based on immunological assumptions compared with interventions based on psychological assumptions. The authors hypothesized that this difference could be explained by higher expectations of patients on interventions assuming physical causation as opposed to interventions assuming psychological causation. Alternatively, they also considered the possibility that systematic differences between immunological and psychological trials, such as illness duration, placebo type, and duration of follow-up, could explain the larger placebo effect in immunological trials. In a meta-regression, however, only intervention type (i.e., psychological or immunological) set out to be significantly associated with a stronger placebo response ($p = .03$), independent from all other factors (8).

Some problems affect the validity and reliability of meta-regression. Primarily, meta-regression is prone to inflate false-positive rates when heterogeneity is present, when there are few studies, and when there are many covariates. Consider the case of two studies producing effect estimates with nonoverlapping CIs: any covariate whose value differs between these studies will be significantly related to the heterogeneity among the studies, and hence, a potential explanation of it, although this explanation could be entirely spurious (58). Furthermore, it is unclear how many covariates can reliably be investigated without the risk of overfitting, and how this depends on the number of studies, the extent of the heterogeneity, and the relative weights of the different studies. Rules of thumb for conventional regression analyses (10–15 observations per covariate assessed, for instance) (56) are not directly relevant to meta-regression, as this type of regression deals with the complexities of heterogeneity and differential study weights. To be on the safe side, meta-analysts who aim to explore heterogeneity using meta-regression should minimize the number of covariates investigated, select those justified through scientific rationale, and specify them in advance (58).

Second, regarding translation of the results of the meta-regression to individuals, the problem of aggregation bias ("ecological fallacy") may arise. This bias refers to the assumption that individuals have the average characteristics of

the group to which they belong and, thus, that relationships observed for groups necessarily hold for individuals. The meta-regression analysis is conducted at the study level and does not include the underlying patient-level variation. The relationship between group means may not reflect the relationship between values of exposure and outcome in an individual, and average quantities instead of person-specific quantities can lead to erroneous conclusions (59). For example, suppose that countries with a high per capita income have high suicide rates. Inferring that increasing personal income at the individual level is also associated with suicide-related mortality can lead to erroneous conclusions, as within countries, suicide-related mortality may be lower in high-income than in low-income persons. Thus, when interesting findings are discovered using meta-regression, person-level data from large cohort studies or trials may be required for confirmation.

Alternatively, meta-analysis of individual patient data (IPD) (also referred to as "mega-analysis") could be considered, in which raw data from every primary study are obtained and transformed to a common format. The strengths of IPD meta-analysis, in general, are that the power is greatly enhanced by the larger number of subjects and that more subgroup analyses can be done. More importantly, however, in observational research, IPD can be used to adjust consistently for confounders. Adjustment for confounders is usually impossible in common meta-analysis, as not all studies perform the same adjustments in their analyses, and they report adjusted analyses in different ways. An example is the meta-analysis on the impact of depression on mortality. In this research, several moderator analyses, such as on measurement instrument to assess depression or duration of follow-up, did not explain heterogeneity (39). Also, the relative risk of mortality was nearly the same in unadjusted and adjusted results, and the amount of heterogeneity was not reduced. Authors argued that one possible explanation for the heterogeneity of the adjusted analyses may be the selection of risk factors, which varied greatly from study to study. One possible solution to this problem would be to pool and reanalyze the original data of all included studies. This can only be done when different studies include comparable measures of the variables to be adjusted for. A major obstacle of IPD meta-analysis is that it is time-consuming and requires cooperation between several research groups, which may not always be attainable. In addition, variables that must be compared will generally be measured using different instruments in the individual studies, and must therefore be harmonized before analysis is possible. Information may be lost during this process of harmonization. A good overview of the methodology of IPD meta-analysis is given by Stewart and Clarke (60).

Interpretation of Meta-Analysis Findings

Sensitivity Analysis

The process of undertaking a meta-analysis inevitably involves many more or less subjective decisions; sensitivity analyses can be conducted to determine whether the assumptions or decisions made have a major effect on the result of the meta-analysis. Thus, sensitivity analysis addresses the question of whether the findings of the meta-analysis are robust to the methods used to obtain them. Examples of sensitivity analyses include assessing the influence of including studies that were doubted to meet eligibility criteria, comparing fixed effect with random effects models, comparing cohort and case-control studies, adding conservative effect size estimations for studies that did not provide adequate data to calculate an effect size, or excluding outlying studies. Two other commonly performed sensitivity analyses are assessing the influence of methodological study quality of primary studies and the influence of publication bias.

Methodological Quality Used in Interpreting Meta-Analysis Results

It remains a matter of debate how the results of quality assessment should be incorporated in the analysis and interpretation of results of meta-analyses. Exploring the effects of quality on the quantitative results by using quality as a weighting factor has been discouraged (61). We recommend using quality scores in a sensitivity analysis, which can demonstrate whether the findings of the meta-analysis are different for low- and high-quality studies. For example, sensitivity analysis in a meta-analysis on the association between positive well-being and mortality in healthy populations indicated a stronger association between positive psychological well-being and reduced mortality in high-quality studies compared with low-quality studies (62). This sensitivity analysis, thus, supports the validity of the overall finding that there is an association.

Publication Bias

Publication bias in observational meta-analyses may lead to inflated effect estimates that tend to be in the hypothesized direction. Several approaches have been developed to assess publication bias. The most well-known approach is the funnel plot—a scatter graph in which, for each primary study, the effect estimate is plotted against a measure of precision (such as sample size, or preferably, the standard error of the effect size) (63). It is expected that more precise studies report effect estimates close to the true effect, whereas effect estimates from less precise studies will scatter more widely. In the absence of publication bias, the plot is expected to resemble a funnel-like shape, which is symmetrical around the summary estimate. In a meta-analysis on decreased cardiac vagal activity in functional somatic syndromes, the funnel plot was not symmetric, as there is a gap where small studies with null findings are expected (Fig. 3*A*). Funnel plots can be visually interpreted, but this is subjective and the agreement between raters and the association between graph ratings and publication bias is found to be poor (64). A test for funnel plot asymmetry formally examines whether the association between estimated effects and a measure of study precision is larger than might be expected to occur by chance. The prin-

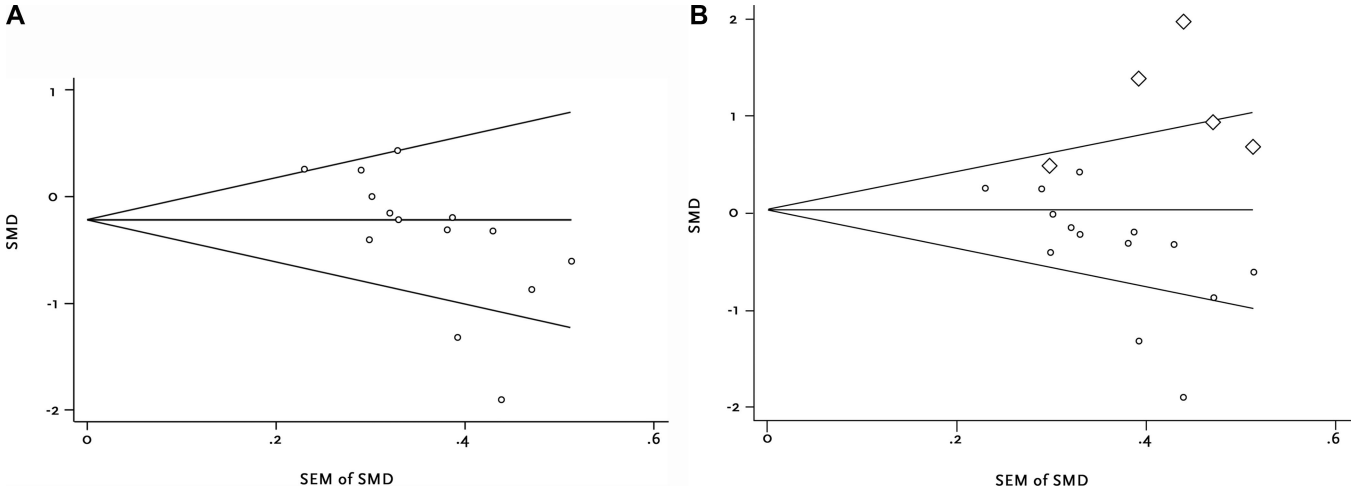

Figure 3. *A*) Funnel plot ($n = 14$) showing the correlation between the standardized mean difference (*SMD*) and its standard error (*SEM*) with pseudo 95% confidence limits. The summary estimate reveals that cardiac vagal activity is significantly lower in patients with functional somatic disorders compared with healthy controls (SMD, −0.32; 95% confidence interval [CI], −0.63 to −0.01; $p = .04$). *B*) Trimmed and filled funnel plot ($n = 19$) showing the correlation between the SMD and its SEM with pseudo 95% CI limits. *Squares* represent the studies that have been filled. The adjusted summary estimate reveals that cardiac vagal activity is not significantly different in patients with functional somatic disorders compared with healthy controls (SMD, 0.01; 95% CI, −0.34 to 0.36; $p = .95$). Reprinted, with minor modifications from Biol Psychol 2009;82:101–110, Tak LM, Riese H, De Bock GH, Manoharan A, Kok IC, Rosmalen JG. As good as it gets? A meta-analysis and systematic review of methodological quality of heart rate variability studies in functional somatic disorders, with permission from Elsevier. (25).

ciple is to relate the effect estimates to their SEM, and to test the null hypothesis that the association is absent. There are many tests for funnel plot asymmetry, which were compared by Rucker and colleagues (65). Two of the most well-known test are the Begg and Mazumdar adjusted rank correlation test (66) and the regression asymmetry test of Egger and colleagues (67). The Begg and Mazumdar test (66) is based on a Kendall's τ rank correlation between the standardized effect size and its SEM. The test of Egger et al. (67) is based on a linear regression of the effect estimate against its standard error, weighted by the inverse of the variance of the effect estimates. However, the tests have low power to detect funnel plot asymmetry and thus do not exclude the presence of publication bias. Publication bias is not the only reason for funnel plot asymmetry. Asymmetry also arises because of small study effects—a tendency for the effects estimated in smaller studies to differ from those in larger studies (68). Small study effects occur because of differences in methodological quality between larger and smaller studies, heterogeneity between studies with different sample sizes (small studies may more likely include selected groups of patients), an effect modifier associated with study precision, or merely chance (67,68). In addition, some effect estimates (e.g., ORs and SMDs) are naturally correlated with their standard errors and can produce spurious asymmetry in a funnel plot (12). Another mathematical estimation of publication bias is provided by the fail safe *N*, which indicates the number of new, unpublished, or unretrieved nonsignificant studies that would be required to lower the significance of a meta-analysis to nonsignificant. A fail safe *N* that is small, particularly compared with the number of studies included in the meta-analysis, indicates that the degree of confidence that can be placed in the main conclusions of the meta-analysis is low. The fail safe *N* has been criticized for two reasons. First, it overemphasizes statistical significance. Second, it is based on the addition of studies that have an average null effect, whereas unpublished studies may also have an effect in the opposite direction as the observed meta-analysis result (69).

An important question is how to proceed when publication bias is suspected. A relatively simple approach to correct for publication bias is the "trim and fill" method (70,71), available in most statistical meta-analysis programs. The principle behind this method is to impute new studies to an asymmetric funnel plot, followed by a meta-analysis that includes the imputed studies. The method works by estimating the number of studies on the right-hand side of the funnel plot that have no counterpart on the left-hand side. Studies causing the asymmetry are then "trimmed" from the right-hand side of the funnel plot, possibly leading to a shift of the reestimated summary estimate that may again create asymmetry. The process is repeated until there is no residual asymmetry, after which the trimmed studies are put back and their missing counterparts are imputed or "filled" by replicating the opposite side of the funnel plot with the mirror axis placed along the adjusted summary estimate. The difference between the original summary estimate and the summary estimate based on the extended data set, including the imputed studies, is assumed to indicate the degree of publication bias. An assumption underlying the trim and fill method is that the magnitude of the effect size, and not the *p* value, determines the chance of publication. Moreover, this technique assumes that publication bias leads to this simple form of funnel plot asymmetry, and that missing effect size estimates are of the same size as those observed in the opposite direction. Never-

theless, it has been shown that the use of the trim and fill method can help to reduce the influence of publication bias on the summary estimates, even though the performance of this method decreases when heterogeneity increases (72). With regard to the studies on cardiac vagal activity in functional somatic syndromes, Egger's test rejected the null hypothesis that there was no funnel plot asymmetry ($p = .01$). The trim and fill method resulted in a fill of five studies and a shift from an initially significant summary estimate to a reestimated summary estimate that was nonsignificant (Fig. 3*B*) (25). This analysis points to the possibility that studies contradicting prevailing beliefs of lower cardiac vagal activity in functional somatic syndromes have not been published. The trim and fill method is recommended to be used as a form of sensitivity analysis of the summary estimate.

Controversy Around Meta-Analysis

There are a number of outspoken critics of meta-analysis. Most points of criticism do not only apply to meta-analysis but also to the entire field of observational research, such as the risk of reporting bias, publication bias, confounding, and lack of comparability between studies. Some even argue that meta-analysis of observational studies should not be done at all, because it would only reinforce the biases inherent to epidemiologic research by creating significant but incorrect results (73). In a properly performed meta-analysis, however, these limitations can be dealt with in a sound way, as has been discussed in this article. Some critics argue that the statistical pooling of data in observational data are highly prone to bias and spurious findings. Instead, it is suggested that it is more important to thoroughly investigate causes of heterogeneity (9). We agree that statistical combination of studies should not generally be the main aim of systematic reviews of observational studies, especially as heterogeneity seems the rule rather than the exception (Table 1). The thorough consideration of possible sources of heterogeneity between studies, by using moderator analysis, meta-regression, and sensitivity analysis should be considered as more important features of meta-analysis in this field. When there are still serious limitations to the results of the meta-analysis, these can be discussed, and interpretation can be adjusted accordingly. Thus, instead of disputing the technique of meta-analysis itself, we feel its undue reputation of providing the final answer should be rectified.

Concluding Remarks

Many papers in *Psychosomatic Medicine* primarily are observational studies aiming to answer etiological questions. Apart from providing a summary estimate, the importance of meta-analyses based on those studies also lies in the identification of sources of bias, heterogeneity, generation of new hypotheses, and the construct of guidelines to conduct better research in the future. Rather than pretending to provide the final, not debatable answer, meta-analysis relies on shared subjectivity. Every analysis inevitably requires certain subjective decisions, but these decisions should be transparent and explicit. The discussion of a meta-analysis should not simply state the results of the statistical pooling, but it should also discuss the level of certainty of the conclusions and any limitations to the interpretation of the findings. Specific guidelines on adequate reporting of meta-analyses based on clinical trials (i.e., preferred reporting items for systematic reviews and meta-analyses [PRISMA]) or on observational studies (i.e., Meta-analysis Of Observational Studies in Epidemiology [MOOSE]) are available (74,75).

This review aimed to demonstrate that performing a meta-analysis is a good way to gain more knowledge concerning a specific research topic. We agree with Rosenthal and DiMatteo (1) when they stated that anyone who is considering a review of the literature has little justification for not doing it quantitatively, as the skills and training required for performing a high-quality meta-analysis are modest. We hope that the number of meta-analyses in *Psychosomatic Medicine* will increase, as they have the ability to produce more knowledge than is provided by the sum of its parts.

REFERENCES

1. Rosenthal R, DiMatteo MR. Meta-analysis: recent developments in quantitative methods for literature reviews. Annu Rev Psychol 2001;52:59–82.
2. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
3. Garg AX, Hackam D, Tonelli M. Systematic review and meta-analysis: when one study is just not enough. Clin J Am Soc Nephrol 2008;3:253–60.
4. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. BMJ 1997;315:1533–7.
5. Freedland KE, Reese RL, Steinmeyer BC. Multivariable models in biobehavioral research. Psychosom Med 2009;71:205–16.
6. Sutton AJ, Higgins JP. Recent developments in meta-analysis. Stat Med 2008;27:625–50.
7. Egger M, Smith GD. Meta-analysis. Potentials and promise. BMJ 1997;315:1371–4.
8. Cho HJ, Hotopf M, Wessely S. The placebo response in the treatment of chronic fatigue syndrome: a systematic review and meta-analysis. Psychosom Med 2005;67:301–13.
9. Egger M, Schneider M, Davey SG. Spurious precision? Meta-analysis of observational studies. BMJ 1998;316:140–4.
10. Rosenthal R. Meta-analysis: a review. Psychosom Med 1991;53:247–71.
11. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997;127:380–7.
12. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1. Hoboken, NJ: Wiley-Blackwell; 2008.
13. Egger M, Zellweger–Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350:326–9.
14. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol 2002;31:115–23.
15. Sampson M, McGowan J. Errors in search strategies were identified by type and frequency. J Clin Epidemiol 2006;59:1057–63.
16. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med 1998;17:841–56.
17. Meade MO, Richardson WS. Selecting and appraising studies for a systematic review. Ann Intern Med 1997;127:531–7.
18. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med 2002;21:1635–40.
19. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666–76.

20. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054–60.
21. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996;313:570–1.
22. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41–4.
23. Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology 2007;18:800–4.
24. Lee W, Bindman J, Ford T, Glozier N, Moran P, Stewart R, Hotopf M. Bias in psychiatric case-control studies: literature survey. Br J Psychiatry 2007;190:204–9.
25. Tak LM, Riese H, De Bock GH, Manoharan A, Kok IC, Rosmalen JG. As good as it gets? A meta-analysis and systematic review of methodological quality of heart rate variability studies in functional somatic disorders. Biol Psychol 2009;82:101–10.
26. DeCoster J. Meta-analysis notes. Available at http://www.stat-help.com/notes.html. Accessed March 4, 2009.
27. Lipsey MW, Wilson DB. Practical Meta-Analysis. Vol 49. Thousand Oaks, CA: Sage Publication; 2001.
28. Dickens C, McGowan L, Dale S. Impact of depression on experimental pain perception: a systematic review of the literature with meta-analysis. Psychosom Med 2003;65:369–75.
29. Greenland S, O'Rourke K. Meta-analysis. In: Rothman KJ, Greenland S, Lash TJ, editors. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
30. Rutledge T, Hogan BE. A quantitative review of prospective evidence linking psychological factors with hypertension development. Psychosom Med 2002;64:758–66.
31. Buckley TC, Kaloupek DG. A meta-analytic examination of basal cardiovascular activity in posttraumatic stress disorder. Psychosom Med 2001;63:585–94.
32. Hedges LV, Olkin I. Statistical methods for meta-analysis. San Diego, CA: Academic Press; 1985.
33. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. J Educ Stat 1981;6:107–28.
34. Cordero DL, Sisto SA, Tapp WN, LaManca JJ, Pareja JG, Natelson BH. Decreased vagal power during treadmill walking in patients with chronic fatigue syndrome. Clin Auton Res 1996;6:329–33.
35. Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, Barrowman NJ. A systematic review identifies a lack of standardization in methods for handling missing variance data. J Clin Epidemiol 2006;59:342–53.
36. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. New York: Academic Press; 1988.
37. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575–600.
38. Van Melle JP, de Jonge P, Spijkerman TA, Tijssen JG, Ormel J, van Veldhuisen DJ, van den Brink RH, van den Berg MP. Prognostic association of depression following myocardial infarction with mortality and cardiovascular events: a meta-analysis. Psychosom Med 2004;66:814–22.
39. Barth J, Schumacher M, Herrmann-Lingen C. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. Psychosom Med 2004;66:802–13.
40. Wulsin LR, Singal BM. Do depressive symptoms increase the risk for the onset of coronary disease? A systematic quantitative review. Psychosom Med 2003;65:201–10.
41. Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000;19:3275–89.
42. Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res 1993;2:121–45.
43. Bax L, Yu LM, Ikeda N, Moons KG. A systematic comparison of software dedicated to meta-analysis of causal studies. BMC Med Res Methodol 2007;7:40.
44. Egger M, Stern JA, Smith GD. Meta-analysis software. BMJ 1998;316:221–5.
45. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. Am J Epidemiol 1995;142:371–82.
46. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.
47. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 1977;35:1–39.
48. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
49. Rosmalen JG, Tak LM, Ormel J, Wessely S, Kok IC, Cleare AJ, Manoharan A. Meta-analysis and meta-regression of HPA axis activity in functional somatic disorders. Abstract 68th Annual Scientific Meeting of the American Psychosomatic Society, Portland 2010. *Psychosom Med* 2010; 72 (abstract).
50. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta-Analysis. 1st ed. West Sussex, UK: Wiley; 2009.
51. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
52. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994;19:309:1351–5.
53. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ 2008;336:1413–5.
54. Berlin JA. Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. Am J Epidemiol 1995;142:383–7.
55. Howren MB, Lamkin DM, Suls J. Associations of depression with C-reactive protein, IL-1, and IL-6: a meta-analysis. Psychosom Med 2009;71:171–86.
56. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004;66:411–21.
57. Meewisse ML, Reitsma JB, de Vries GJ, Gersons BP, Olff M. Cortisol and post-traumatic stress disorder in adults: systematic review and meta-analysis. Br J Psychiatry 2007;191:387–92.
58. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004;23:1663–82.
59. Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. Am J Public Health 1994;84:819–24.
60. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. Stat Med 1995;14:2057–79.
61. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992;45:255–65.
62. Chida Y, Steptoe A. Positive psychological well-being and mortality: a quantitative review of prospective observational studies. Psychosom Med 2008;70:741–56.
63. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001;54:1046–55.
64. Bax L, Ikeda N, Fukui N, Yaju Y, Tsuruta H, Moons KG. More than numbers: the power of graphs in meta-analysis. Am J Epidemiol 2009;169:249–55.
65. Rucker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. Stat Med 2008;27:746–63.
66. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088–101.
67. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34.
68. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 2000;53:1119–29.
69. Evans S. Statistician's comment on: misleading meta-analysis. "Fail safe N" is a useful mathematical measure of the stability of results. BMJ 1996;312:125.
70. Duval S, Tweedie R. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. J Am Stat Assoc 2000;95:89–98.
71. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000;56:455–63.
72. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. Stat Med 2007;26:4544–62.
73. Shapiro S. Meta-analysis/Shmeta-analysis. Am J Epidemiol 1994;140:771–8.

META-ANALYSIS IN PSYCHOSOMATIC MEDICINE

74. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
75. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000; 283:2008–12.
76. Chida Y, Hamer M, Steptoe A. A bidirectional relationship between psychosocial factors and atopic disorders: a systematic review and meta-analysis. Psychosom Med 2008;70:102–16.
77. Steffen PR, Smith TB, Larson M, Butler L. Acculturation to Western society as a risk factor for high blood pressure: a meta-analytic review. Psychosom Med 2006;68:386–97.
78. Carter FA, Frampton CM, Mulder RT. Cesarean section and postpartum depression: a review of the evidence examining the link. Psychosom Med 2006;68:321–30.
79. Sundin EC, Horowitz MJ. Horowitz's Impact of Event Scale evaluation of 20 years of use. Psychosom Med 2003;65:870–6.
80. Henningsen P, Zimmermann T, Sattel H. Medically unexplained physical symptoms, anxiety, and depression: a meta-analytic review. Psychosom Med 2003;65:528–33.
81. Dickens C, McGowan L, Clark–Carter D, Creed F. Depression in rheumatoid arthritis: a systematic review of the literature with meta-analysis. Psychosom Med 2002;64:52–60.
82. De Groot M, Anderson R, Freedland KE, Clouse RE, Lustman PJ. Association of depression and diabetes complications: a meta-analysis. Psychosom Med 2001;63:619–30.